概率论与数理统计知识概览

——

适用于考研、复习巩固

Dezeming Family

2021 年 7 月 20 日

DezemingFamily 系列书和小册子因为是电子书，所以可以很方便地进行修改和重新发布。如果您获得了 DezemingFamily 的系列书，可以从我们的网站 [https://dezeming.top/] 找到最新版。对书的内容建议和出现的错误欢迎在网站留言。

20210722：完成第一版。

目录

笔者在本科保研之后，曾经因为兴趣，兼职过一段时间的家教，主要讲线性代数和概率论的内容。我在讲解概率论和线代中，发现很多同学都不知道怎么去复习和掌握这些基本数学科目的纲要，导致复习也遇到很多困难，因此我打算将大学本科的概率论知识（非数学类专业）的内容，**用讲故事的方式**进行一下知识串讲和总结。

这本小册子只会有一章，将概率论作为一个整体，我相信在阅读了这本小册子以后，大家就能对概率论到底讲了什么、应该怎么去学习有一个全面和深刻的认识，再去复习知识点也会变得得心应手。那么，现在我们就开始讲述这个故事——概率论与数理统计。

一 概率论与数理统计

现在，让我们合上课本，思考一下，概率论究竟讲了些什么呢？

你抛一枚硬币，它会出现正面和反面；你投掷一个骰子，它正面朝上的数字可能有六种，这些，都是最基本的概率知识了。

但是，现在情况变得复杂了，你已经投掷了一枚硬币，正面朝上，你再投掷一次，它正面朝上的概率是多少呢？我们单独思考每一次的投掷，它出现正面朝上的概率都是 $\frac{1}{2}$，但是，当你把两次投掷考虑为一个整体，那么，就要将这两种情况都考虑进来，因此，条件概率产生了。条件概率是指，当我们得知已经发生了某件事，我们想知道在这个前提下发生另一件事的概率。

伴随着条件概率，从而又出现了两个相对不好理解的公式——贝叶斯公式和全概率公式：当我们需要求得已经发生某件事的前提下，发生另一件事的概率，那么我们就会需要这两个公式，比如你们班有 50 个人，其中 40 个人考试及格了，且 30 个人高于 80 分，那么如果你随机抽取一个人，那个人你知道他已经及格了，那么他高于 80 分的概率是多少呢？这就可以用这些公式来进行计算。

有些事件并不是互相关联的，比如南极有只企鹅被海象吃掉这件事，跟你闺蜜考试不及格应该是没有什么联系的。如果发生两件事之间互不影响，我们就认为这两件事互相独立，比如，投掷两次骰子，正面朝上的数字都是 6 的可能性是 $\frac{1}{36}$，它等于两次投掷分别出现 6 的概率相乘，即 $\frac{1}{6}\times\frac{1}{6}$。独立事件和条件概率是比较难区分和容易混淆的情况，但其实多做几个题，就能很容易把它们的应用情景区分开。

直到现在，我们还是在离散的场景中研究概率问题，但有些时候，有些数据并不是离散的，例如某地区一年内降水量，恰好是 723 毫米的概率为 0，因为几乎不可能恰好就是这个值嘛！我们某次观察到的降水量是 721.3 毫米，但降水量恰好是 721.3 毫米的概率却也是 0，这是因为对于连续分布的数据，我们无法直接描述单独某一点的概率。

尽管无法描述在某一点的概率，但我们可以描述在某个区间的概率，比如，降水量在 700-720 毫米的概率为 11%，我们逐步缩小这个区间，就像求导一样，就能得到概率密度函数 PDF，我们对 PDF 在某个区间进行积分，得到的就是落在这个区间的概率，又叫做累积概率密度函数 CDF。连续的情况和离散的情况并没有太大的不同，只是连续的情况可以引出很多典型的概率分布，例如高斯分布。

现在，当我们已经知道了什么是概率，什么是概率分布，我们就需要研究一些对应的数字特征，什么特征呢？比如，我们统计了一个班的同学的身高，我们想知道自己算高的还是矮的，因此我们就通常求出身高的平均数。在数理统计中，平均值有专业的名词描述，即期望。这里的期望也不止离散数据的期望，对于连续数据，我们仍然可以求出期望，比如我们知道了降水量的分布规律，我们就可以求出降水量的期望值。

除了期望，度量统计数据的是否更集中的量——方差也是很重要的数字特征，例如，你们班同学身高的方差较小，说明大家身高都比较集中；较大则说明你们班有的人很高，有的人很矮，分布不那么集中。

有时候我们统计量不止有一个，会有多个，比如身高和体重，我们可以使用协方差来描述身高和体重的分布关系，以及使用相关系数来描述身高和体重之间的联系。

我们有些时候并不能有足够的精力和能力去统计所有的数据，而只能抽取一些样本来进行估计。例如要统计中国所有中学生的平均身高，我们可能会在每个省份抽取一些中学生，然后作为样本，来估计总体。样本估计得到的期望值会随着样本数量的增加而越来越准确，这个“准确度”应该怎么去衡量呢？这就用到了大数定律和中心极限定律，这两个理论告诉了我们，使用样本估计得到的期望值会满足什么规律。

样本不但能够来估计总体的期望，还能估计总体的方差。这些估计方法需要尽可能准确，如何保证尽可能准确呢？我们希望估计的样本方差的期望值等于总体的方差，也就是说，我们随机检测 100 次，每次抽取不同的随机独立样本，然后得到 100 个对总体方差的估计，我们希望这些估计的期望值等于总体方差值，因此，就演化出了一些理论和方法。同理，样本协方差、相关系数也会被用来估计总体的协方差和相关系数。

当我们有一些样本时，我们希望用样本来估计总体的一些参数，那么我们就需要进行参数估计，例如当我们知道样本符合正态分布时，我们希望使用样本来估计正态分布的期望和方差，这就构成了参数点估计。有些时候我们不需要那么精确的估计，而是希望有一个区间波动范围，因此就需要进行参数区间估计。

我们不能只进行估计，还需要判断这些估计方法的准确性，因此，一些估计效果的评价指标也需要考虑，例如无偏性估计、有效性估计等。

概率论的进阶是随机过程，或者统计学习，简单的统计学习，例如线性回归分析，这些内容就稍微偏向实际应用了，但万变不离其宗，它描述的也是总结样本规律，获得模型的过程。

至此，概率论的基础知识也就介绍完了。至于它的进阶内容，也总离不开方差、期望等基础概念。中心极限定律、大数定律等内容也是我们常常会用到去证明其他估计器收敛性的重要方法。大家可以根据我的讲解步骤，系统地体会和感受一下概率论这门学科，当你能够将这整个体系联系到一起的时候，你对概率论的理解就会变得深刻。

参考文献

[1] Nothing